



Deccan Education Society's  
FERGUSSON COLLEGE (AUTONOMOUS), PUNE – 4

**Department of Computer Science**

**A**  
**Project Report**  
**On**  
**Matchmaking WebApp**

**In partial fulfillment of requirements of the completion of**  
**M.Sc. (DS) - II**  
**Semester – III**

**SUBMITTED BY:**  
**Siddhant Baliram Fulzele (Roll no. 11114)**

**Under the Guidance of**  
**Mrs. Swati Satpute**  
**(2020 – 2021)**

# CERTIFICATE

This is to certify that the project entitled **Matchmaking WebApp** submitted by

1. \_\_\_\_\_
2. \_\_\_\_\_
3. \_\_\_\_\_

in partial fulfillment of the requirement of the completion of M.Sc. (Data Science)-II [Semester-III], has been carried out by them under our guidance satisfactorily during the academic year 2020-2021.

Place: Pune

Date:    /    /2021

**Head of Department**  
**Department of Computer Science**  
**Fergusson College, Pune**

**Project Guide:**

1. \_\_\_\_\_

**Examiners Name**

**Sign**

- |          |       |
|----------|-------|
| 1. _____ | _____ |
| 2. _____ | _____ |

## ACKNOWLEDGEMENTS

We would sincerely like to thank our guide **Mrs. Swati Satpute** for her support, sincere guidance, timely help and valuable suggestions without which our project would have been impossible. We would also like to express our deepest gratitude to Computer Science Department **HOD Dr. Kavita Khobragade** ma'am and all those who have directly or indirectly helped us in completing this project.

Siddhant Fulzele 11114

1. INTRODUCTION .....	5
1.1. PROJECT DESCRIPTION .....	5
1.2. PROBLEM STATEMENT .....	5
2. DATA .....	6
2.1. DATA SOURCE .....	6
2.2. DATA DESCRIPTION .....	6
2.3. DATA FLOW .....	7
2.4. DATA PREPROCESSING .....	8
3. VISUAL EXPLORATION .....	11
3.1. TYPES OF VISUALIZATION .....	11
3.2. TECHNOLOGIES USED FOR VISUALIZATION .....	11
4. MODEL BUILDING .....	12
4.1. OBJECTIVE .....	12
4.2. ALGORITHMS USED .....	12
4.3. PERFORMANCE METRICS .....	13
5. SCREENSHOTS .....	15
6. ANALYSIS .....	19
7. FUTURE ENHANCEMENT .....	20
8. BIBLIOGRAPHY .....	21

# **1. INTRODUCTION**

## **1.1. PROJECT DESCRIPTION**

Nowadays there are a lot of dating apps to help people to find their significant others. Some dating apps are quite popular among young people also there are some dating apps especially for the elderly. The purpose of those apps is to filter potential matches based on the users' personal preferences, such as height, degree, habit, region, and occupation, etc.

Dating apps have been beneficial for many people who have fallen in love through those apps. But many people have bad experience regarding the matches on the app etc. So, using machine learning we can enhance the dating app experience.

## **1.2. PROBLEM STATEMENT**

With the help of Machine learning, we can improve the overall experience of dating/matchmaking apps to reduce the amount of time spent swiping. We could use AI to find users that are like each other and recommend dating profiles the same way Netflix recommends different movies. These AI systems like Netflix, Amazon etc. have been used in past. Similarly, we will use such AI systems for our dating app.

This type of dating app will make the tedious process of finding a partner online with fun. With machine learning, profiles can be potentially be clustered together with other similar profiles. This will reduce the time-consuming process of going through the non-compatible profiles.

## 2. DATA

### 2.1. DATA SOURCE

Data was taken from the following source:

<https://www.hackerearth.com/challenges/competitive/hackerearth-machine-learning-challenge-predict-match-percentage/>

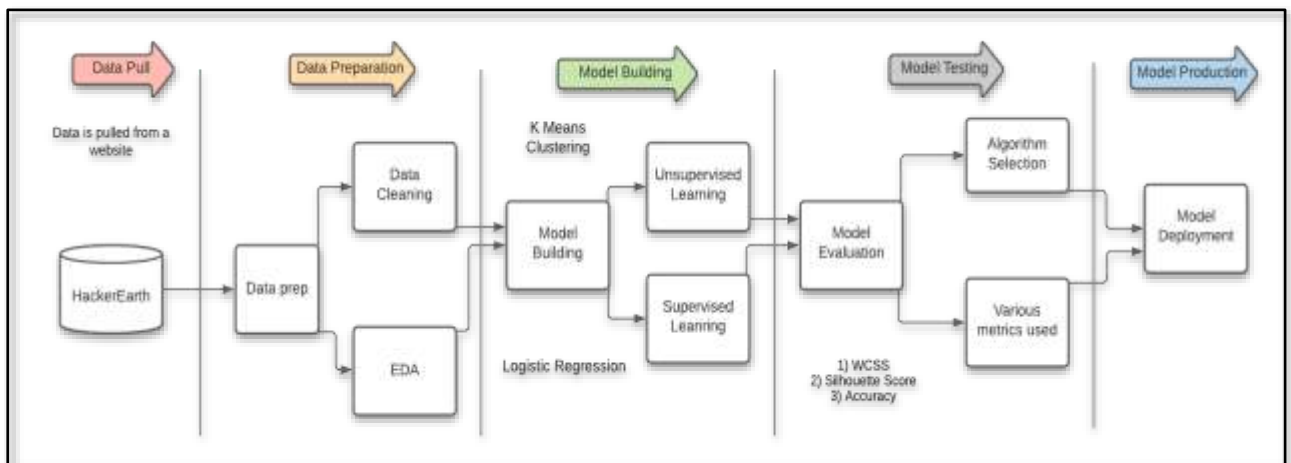
### 2.2. DATA DESCRIPTION

Column Name	Description
User ID	Specifies the number/login id of the user
Username	Specifies the username
Age	Specifies the Age of the user
Status	Specifies the relationship status
Sex	Specifies the sex of the user
Orientation	Tells us whether the user is straight, bisexual or gay.
Drinks	Specifies the frequency of users' drinking
Drugs	Specifies whether the user is having drugs
Height	Specifies the height of the user
Job	Specifies the users' job
Location	Specifies the location where the user is currently staying.
Pets	Specifies whether the user likes pets or not.
Smokes	Specifies whether the user smokes or not
Body profile	Specifies the body profile of the user (fit, athletic etc.)
Education Level	Specifies the users' education.

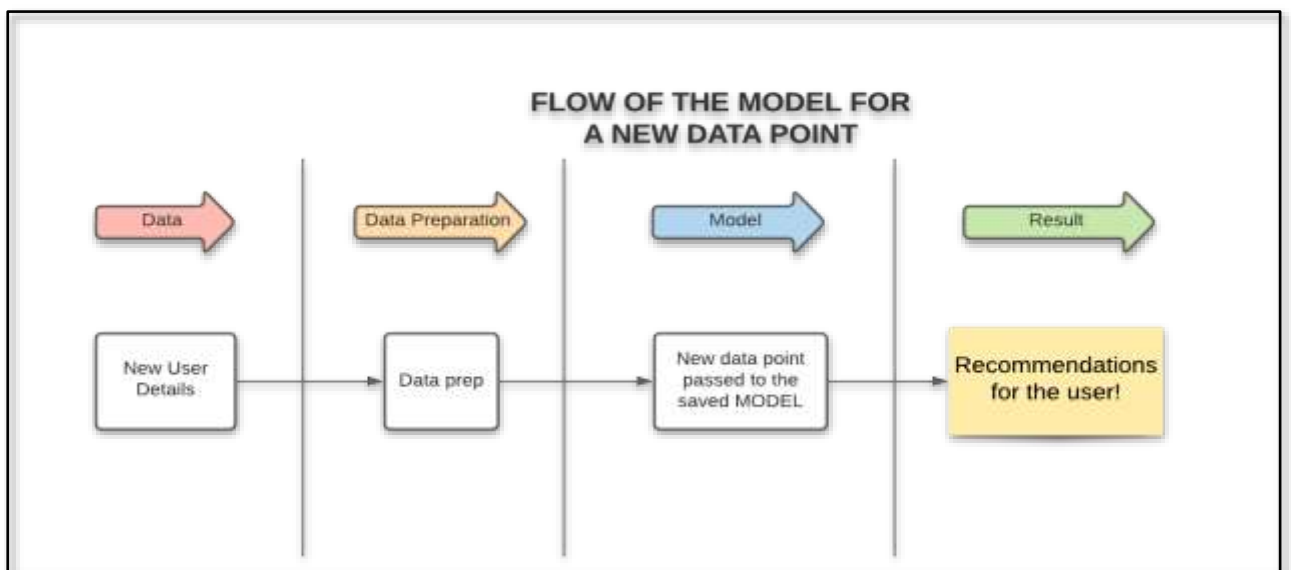
Dropped out	Specifies whether the user has dropped out of the college
Bio	Specifies short information about the user, generally hobbies, nature of that user etc.
Interests	Specifies the interests of the user.
Location Preference	Specifies the location preference of the user

## 2.3. DATA FLOW

Data Life Cycle



New Data Point Life Cycle



## 2.4. DATA PREPROCESSING

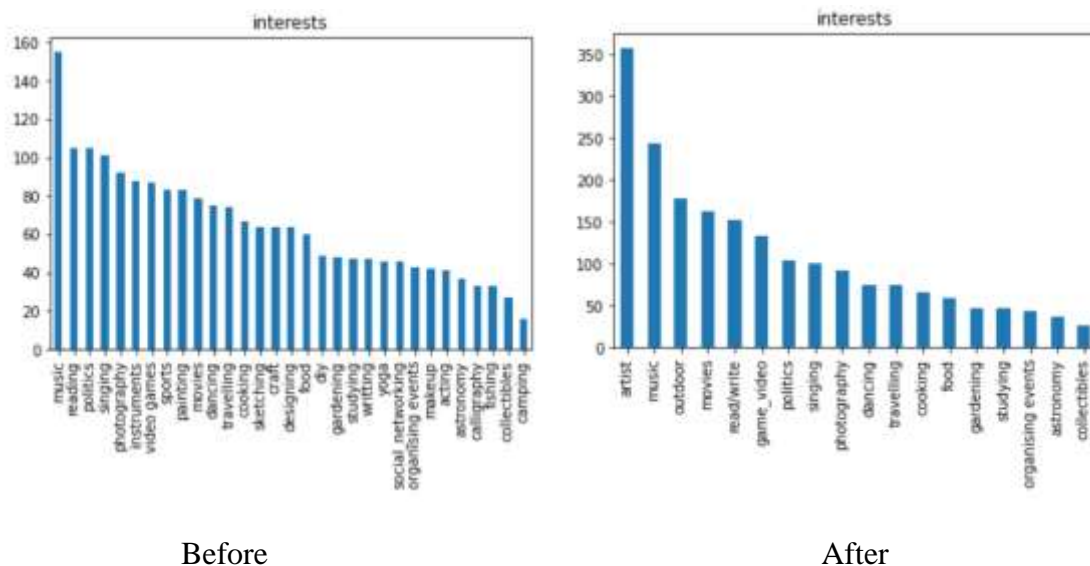
Once the basic data cleaning, importing libraries is done we start clustering of data. Each profile belongs to a specific cluster number of groups. We successfully obtained the clustered groups into the data frame. With the clustered profile data, we sorted profiles based on the similarity. Clustering helps us to answer questions and find more interesting insights.

Natural Language Processing (NLP) – For some columns like bios or hobbies, there is a need to extract useful information from textual data. We can use NLP based on this textual data and combine them with algorithms to get more accurate matches. NLP can also be useful in different aspects such as – trending groups of people, top 10 matches etc.

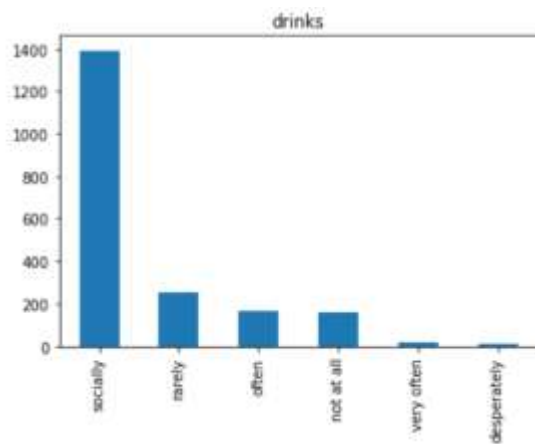
The recommendations system is good for matchmaking. Not only it can recommend good matches according to your basic preferences but also, we can analyze their user behavior.

### PREPROCESSING

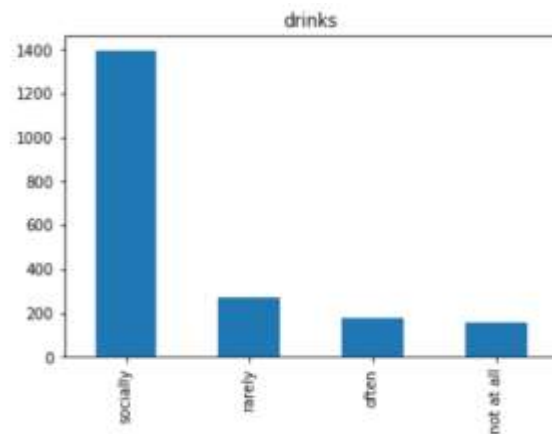
The data has too many categorical columns to handle. For actual model building, a smaller number of columns are used. The main concern was to reduce the number of columns.







Before



After

As shown above we tried to reduce the number of distinct values present in all columns by using count plot (we kept values with high frequency count) and we merged similar categories using domain knowledge (like for drinks column we merged very often and desperately to often).

All this because in the end we were going to apply one hot encoding to the data so blindly applying one hot encoding to the data would create very large number of features so to avoid that the preprocessing was done in above manner.

## NLP PREPROCESSING

Load the necessary libraries as well as libraries from NLTK. The NLTK library will be the primary library we use to process the text data from the dating profile bios. Following are the preprocessing steps:

- A. **Tokenizing the user bios** - Tokenization is the process of splitting up sentences into individual tokens. A function is created for tokenization, in order to have cleaned bio. Within the function -
  - a. Create a library of stop words (the, a, of etc.) to exclude these words which do not make any sense.
  - b. Lowercase the entire string or sentence
  - c. Replace the punctuations.
  - d. Split the string or sentence on the spaces between each word.
  - e. Iterate through each word in the sentence then lemmatize the word or remove if it is a stop word.
- B. **Lemmatizing** – Breaking down the words to their base form of the word, lemmatization is the best option.
- C. **Bigrams** - implement some *bigrams* (pairs of words). We create a list of pairs of words and their respective frequency scores.

'Since start I knew you wanted to talk something, there was some burden like thing you are carrying within you. You say freely I would listen to it, don't hold any hesitations, trust me.'

Sample of user's bio value in raw data

'Knowing each other'

After NLP modelling that would be converted into this label.

Likewise, we transformed all the bio values into 6 relevant labels and applied one hot encoding on that too.

# VISUAL EXPLORATION

## 2.5. TYPES OF VISUALIZATION

- A. **BAR PLOT** – This graph is used for studying the count of the age, height, education level i.e. the numeric columns as well as for the categorical columns.

For NLP, the most frequent words, unique words and their count is plotted using bar plot.

- B. **BOX PLOT** - This graph is used for detecting any outliers in the dataset.
- C. **LINE GRAPH** – For looking at the model performance, line graph is used.
- D. **BUBBLE CHART** - Bubble Charts are typically used to compare and show the relationships between categorized circles, by the use of positioning and proportions. The overall picture of Bubble Charts can be used to analyze for patterns/correlations.

## 2.6. TECHNOLOGIES USED FOR VISUALIZATION

Python libraries like, Matplotlib, Seaborn are used for visualization.

For dashboard, Tableau is used.

## 3. MODEL BUILDING

### 3.1. OBJECTIVE

Clustering is used to cluster the dating profiles with one another. By doing so, the users will get the matches like themselves, instead of the profiles unlike their own. With machine learning, profiles can potentially be clustered together with other similar profiles. This will reduce the number of profiles that are not compatible with one another. From these clusters, users can find other users more like them.

### 3.2. ALGORITHMS USED

The dating app is built using algorithms of Machine Learning. More specifically, utilizing unsupervised machine learning in the form of clustering is the main objective. Algorithms used are:

#### UNSUPERVISED

**K-Means Clustering** – This is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. You'll define a target number  $k$ , which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster. The K-means algorithm identifies  $k$  number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. The '*means*' in the K-means refers to averaging of the data; that is, finding the centroid.

#### SUPERVISED

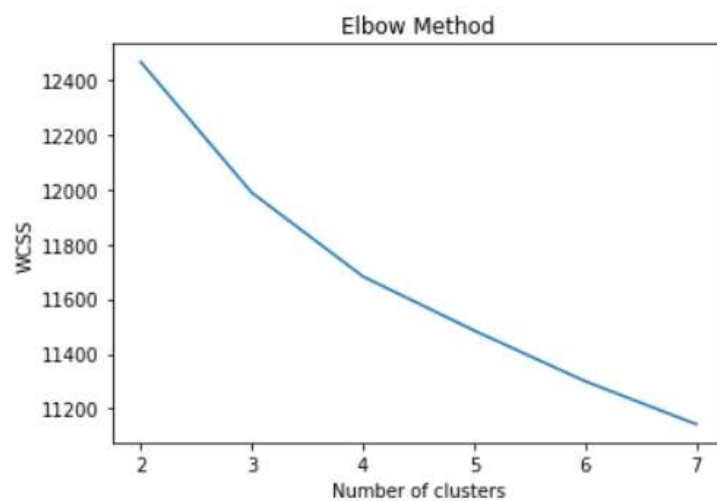
**Logistic Regression** – This is a classification algorithm used to assign observations to a discrete set of classes. It is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends to limit the cost function between 0 and 1.

### 3.3. PERFORMANCE METRICS

For various algorithms, depending on supervised and unsupervised learning different performance metrics are used.

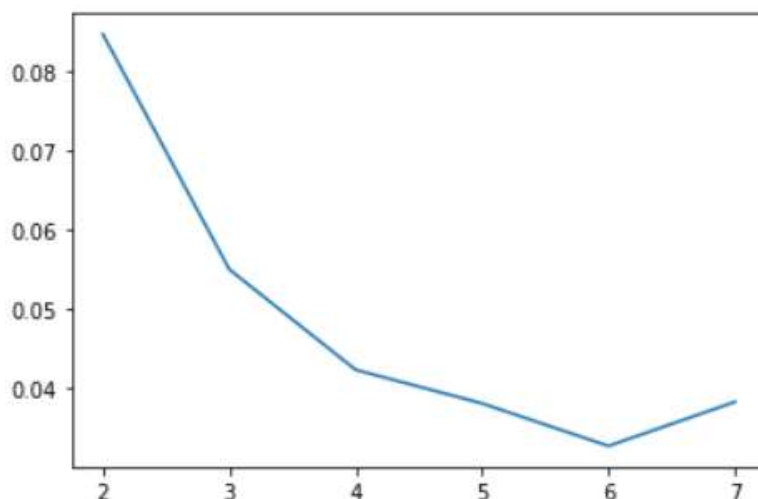
#### UNSUPERVISED

##### WCSS –

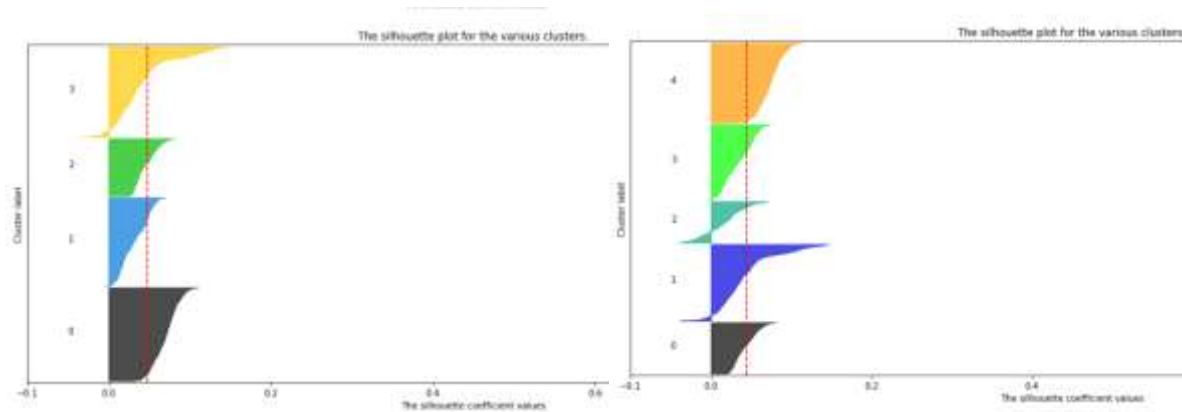


There is slight edge for number of clusters 4. So we decided to analyze performance of 4 and 5 clusters.

##### Silhouette Score (Higher the better) -



Silhouette score of number 4 is higher than 5 so it tells number of clusters should be 4.



Clusters in left side plot looks well distributed and leakage(color spill towards left side) is also less and in addition to this volume of each color(which represent number of users in that cluster) looks well distribute in left plot. So this confirms the number of clusters to be 4.

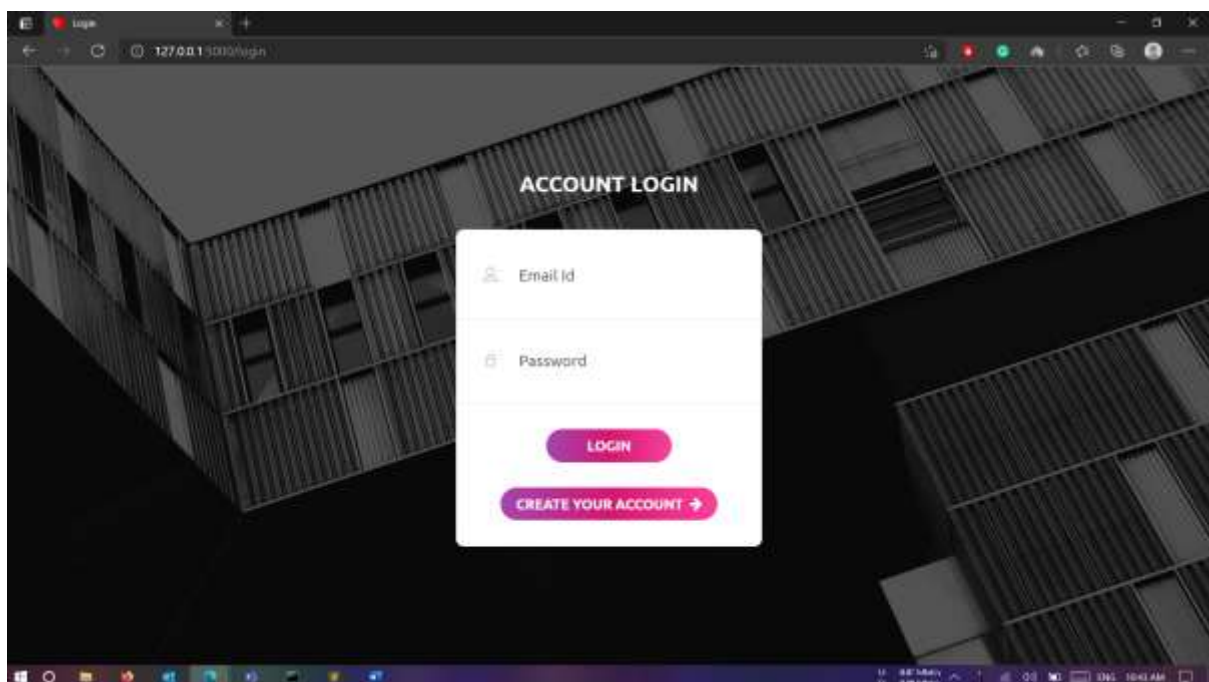
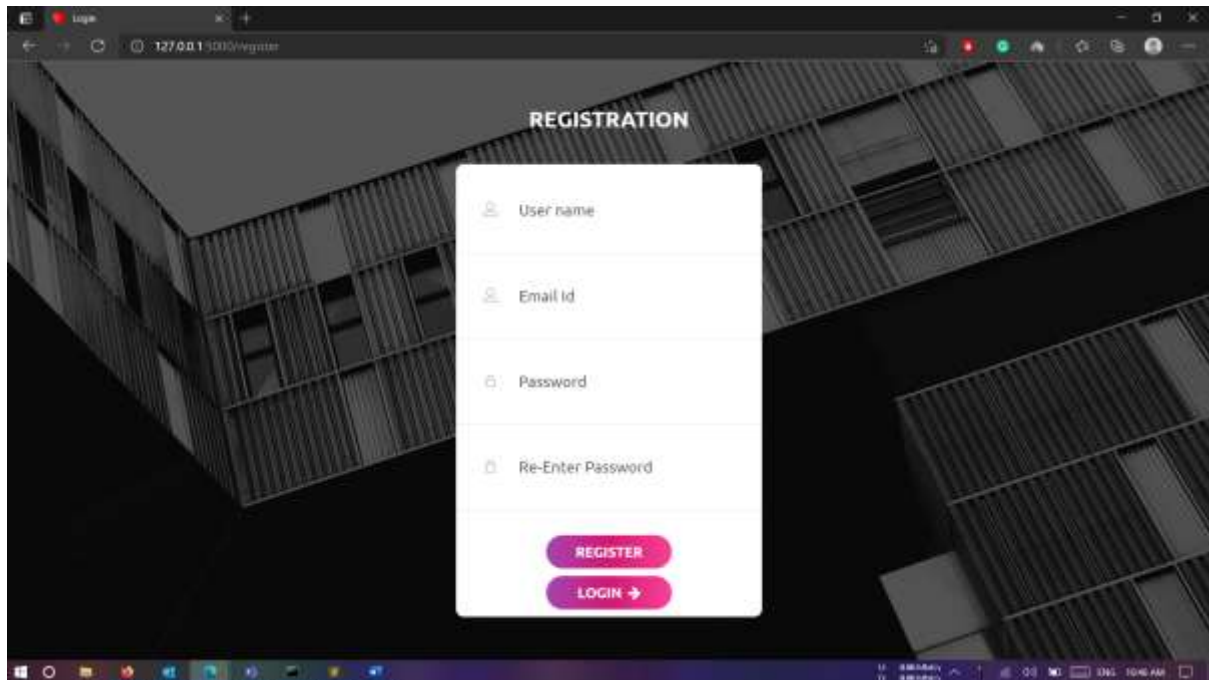
## SUPERVISED

### Confusion Matrix –

```
[[172    0    0    0]
 [  0  156    0    0]
 [  0    0  103    0]
 [  0    0    0  170]]
```

Above is the confusion matrix of logistic model's output on test split. We can see 0 misclassification this is due to the kind of data we provided to model was having values either 0 or 1. So the data was linearly separable, and model could find it easy to classify samples as no continues values were present in data.

## 4. SCREENSHOTS



127.0.0.1:3000/completeProfile

### SIDDFULZELE, WE LOVE TO KNOW ABOUT YOU..!

BE PATIENT..!! THE MORE INFO YOU PROVIDE , BETTER THE RECOMMENDATIONS WE GIVE.

#### Complete Your Profile...


Age : <input type="text" value="Age"/>	Sex : <input type="radio"/> Male <input type="radio"/> Female	Height : <input type="text" value="Height in inches"/>	Drinks : <input type="text" value="Not at All"/>
Orientation : <input type="text" value="Straight"/>	Status : <input type="text" value="Single"/>	Smokes : <input type="text" value="Yes"/>	Drugs : <input type="text" value="Never"/>
Education Level : <input type="text" value="1"/>	Dropped Out : <input type="text" value="No"/>	Job : <input type="text" value="science / tech / engineering"/>	Interests : <input type="text" value="Like : Sports / Instruments / Writing"/>
Body Profile : <input type="text" value="Fit"/>	Pets : <input type="text" value="Like : likes dogs / likes cats / dislikes dogs"/>	Location : <input type="text" value="san francisco"/>	Location Preference : <input type="text" value="anywhere"/>

Bio :

[NEXT →](#)

[CANCEL ←](#)

127.0.0.1:3000/userHome



**It's Siddfulzele**


Age : 22      Gender : Male

Status : Single      Location : San Francisco

Orientation : Straight      Height : 52 inch

[Sign Out](#)


**Kimberly Wheaton**



Sex : F   Age : 26  
From : **Berkeley, California**

[Dislike](#) [Like](#)


**Concepcion Love**



Sex : F   Age : 26  
From : **Berkeley, California**

[Dislike](#) [Like](#)


**Lillian Magana**



Sex : F   Age : 27  
From : **San Francisco, California**

[Dislike](#) [Like](#)


**Lynn Quick**



Sex : F   Age : 26  
From : **San Francisco, California**

[Dislike](#) [Like](#)


**Sabrina Richardson**



Sex : F   Age : 25  
From : **Oakland, California**

[Dislike](#) [Like](#)

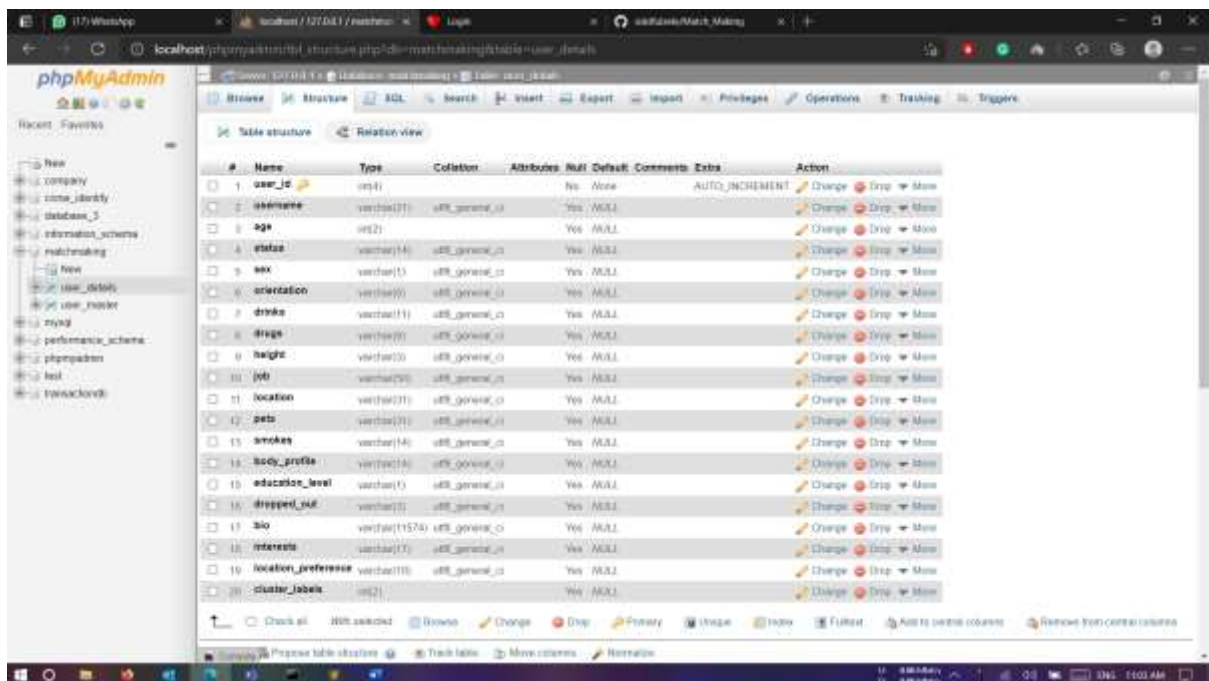
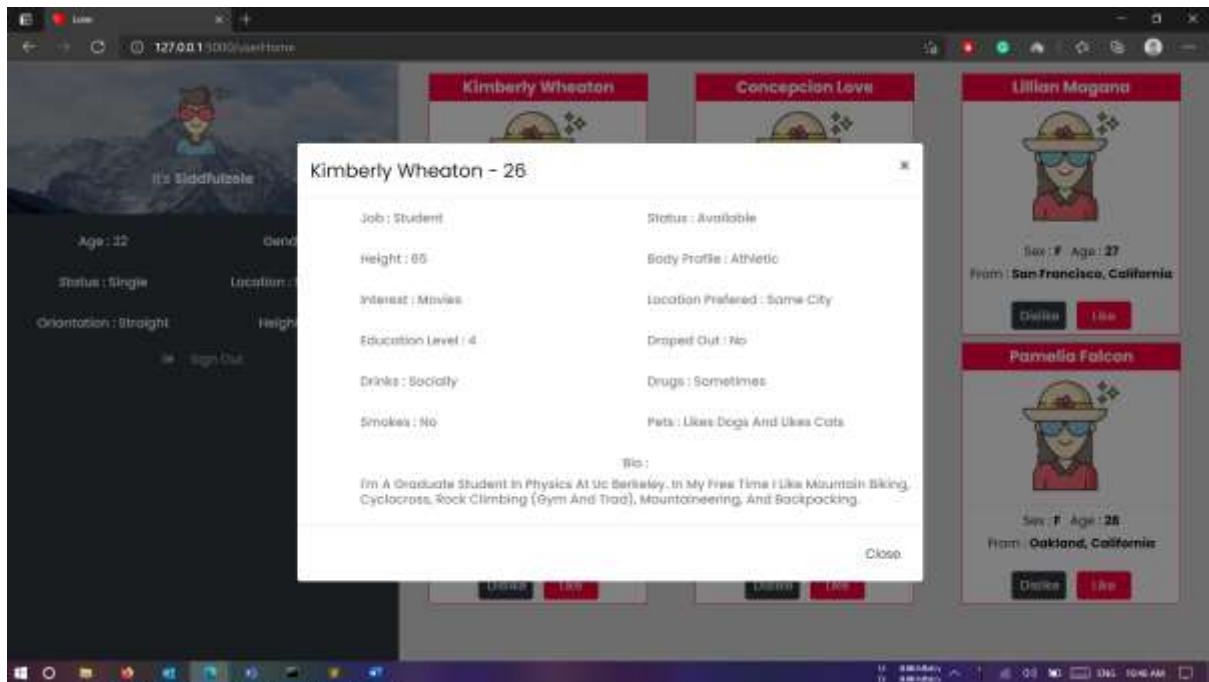
**Pamella Falcon**



Sex : F   Age : 28  
From : **Oakland, California**

[Dislike](#) [Like](#)





localhost/phpmyadmin/structure.php?db=matchmaking&table=user\_master

phpMyAdmin

Recent Favorites

- New
- company
- code\_identity
- database\_5
- information\_schema
- matchmaking
  - New
  - user\_detail
  - user\_master
- mysql
- performance\_schema
- phpmyadmin
- test
- testcockonb

Structure

Table structure

#	Name	Type	Collation	Attributes	Null	Default	Comments	Extra	Action
1	user_id	int(4)			No	None		AUTO_INCREMENT	Change Drop Move
2	username	varchar(31) utf8_general_ci			Yes	NULL			Change Drop Move
3	email	varchar(33) utf8_general_ci			Yes	NULL			Change Drop Move
4	password	varchar(32) utf8_general_ci			Yes	NULL			Change Drop Move
5	isProfileCompleted	varchar(1) utf8_general_ci			Yes	NULL			Change Drop Move

Check all With selected Show Change Drop Primary Unique Index Fulltext Add to default columns Remove from default columns

Print Properties table structure Track table Move column Rename

Add 1 column(s) after isProfileCompleted Go

Indexes

Action	KeyName	Type	Unique	Indexed	Columns	Cardinality	Collation	Null	Comment
Edit Drop	PRIMARY	MYISAM	Yes	No	user_id	2038	A	No	

Create an index on 1 column(s) Go

Partitions

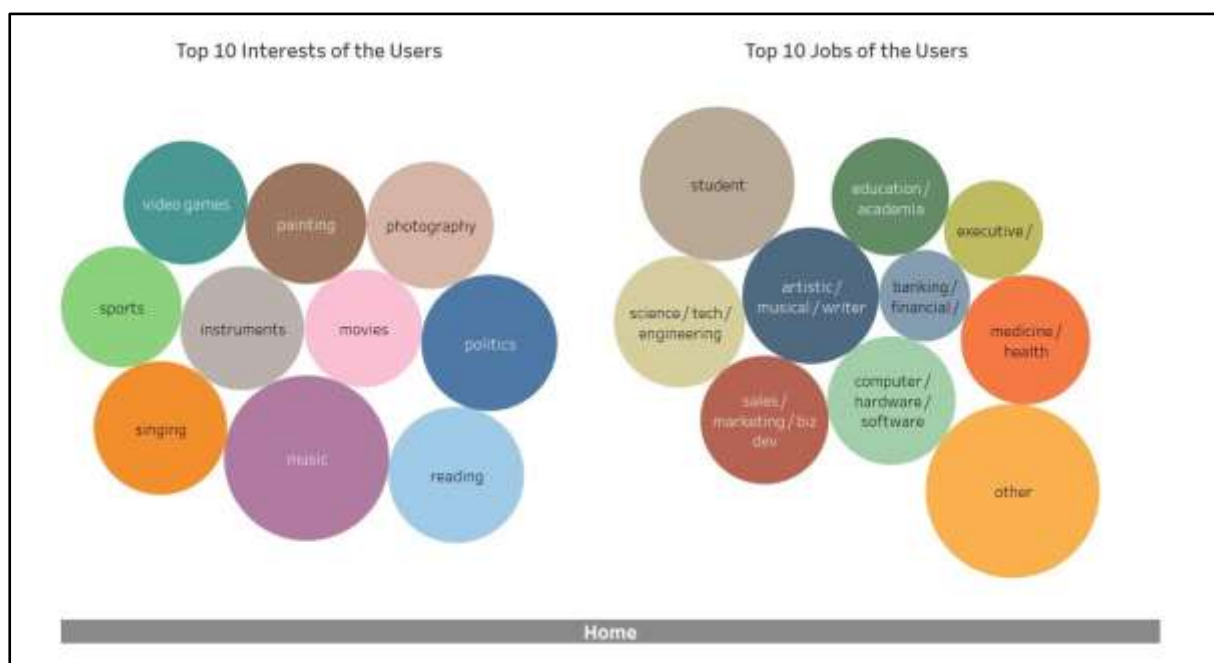
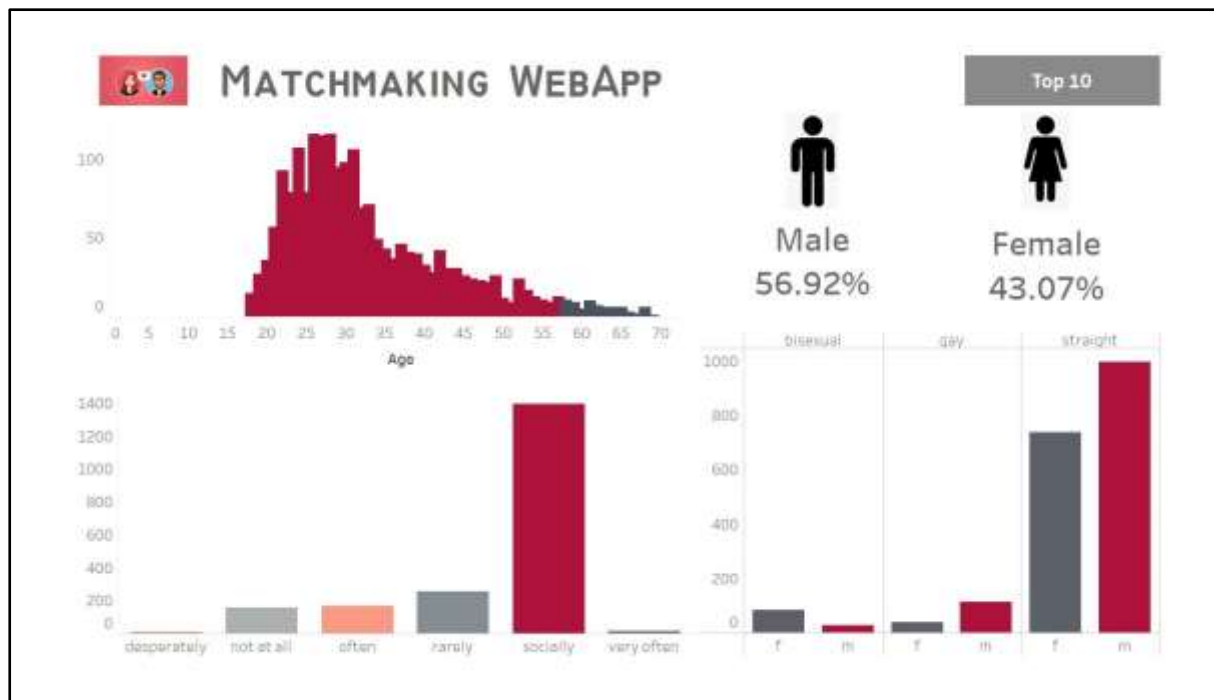
No partitioning defined.

Partition table

MySQL 5.6.23

17:00:00 17/04/2017

## 5. ANALYSIS



Published on Tableau Public -

[https://public.tableau.com/app/profile/trupti5700/viz/MatchMaking\\_Dashboard/Dashboard1](https://public.tableau.com/app/profile/trupti5700/viz/MatchMaking_Dashboard/Dashboard1)

## **6. FUTURE ENHANCEMENT**

To conclude, **Matchmaking WebApp** overcomes many limitations. Following are the advantages of this app like,

- a. Easy implementation environment.
- b. Spend less time swiping
- c. Recommendations based on personal interests

### **Scope for future development**

There are other potential improvements to be made in this project. Following are the future scope for the project.

- a. Implementing a way to include a chat facility to the recommended users.
- b. Users will also be able to upload their own photo, instead of the animations provided right now.
- c. Users can also link social media accounts to their profiles.

## 7. BIBLIOGRAPHY

- d. Sklearn Documentation for Algorithms - [https://scikit-learn.org/stable/unsupervised\\_learning.html](https://scikit-learn.org/stable/unsupervised_learning.html)
- e. Documentations for Visualization
  - i. Matplotlib - <https://matplotlib.org/>
  - ii. Seaborn - <https://seaborn.pydata.org/>
  - iii. Word Cloud for NLP - <https://pypi.org/project/wordcloud/>
- f. Pandas Documentation - [https://pandas.pydata.org/docs/user\\_guide/index.html](https://pandas.pydata.org/docs/user_guide/index.html)